архив журнала "Звукорежиссер" : 2002 : №4

Часть 17.3

Слух и речь, ч. 3

Акустические характеристики речи

Ирина Алдошина

Речевой сигнал имеет двойственную природу – с одной стороны, это обычный акустический сигнал, который представляет собой процесс распространения энергии акустических колебаний в упругой среде. Как любой акустический сигнал, он может быть представлен в виде звуковых волн, представляющих собой распространение процессов сжатия и разряжения частиц среды, формы фронтов которых зависят от свойств источника и условий распространения. Поэтому, как и другие акустические сигналы, речь характеризуется определенным набором объективных характеристик: зависимостью звукового давления от времени (временной структурой звуковой волны), длительностью звучания, спектральным составом, местом расположения источника в пространстве и пр.

С другой стороны, речь как физическое явление вызывает определенные субъективные слуховые ощущения (громкости, высоты, тембра, локализации, маскировки и др.), именно проблемам их взаимодействия и были посвящены предыдущие статьи по психоакустике.

Речевой сигнал подвергается такой же процедуре обработки в слуховой системе, как и любой другой акустический сигнал, т. е. на основе его анализа формируются те же слуховые ощущения – например, восприятие речи на абсолютно незнакомом языке ничем не отличается от восприятия окружающей акустической информации – шума, свиста, щелчков и др. Однако если человек воспринимает речь на языке, которому он был предварительно обучен, то наряду с обработкой чисто акустической информации (громкости, высоты, тембра и пр.) происходит фонетическая, а вслед за ней и семантическая расшифровка информации, для чего подключаются специальные отделы головного мозга.

На протяжении уже многих десятилетий, и особенно интенсивно в последние годы, в связи с развитием технологии и систем автоматического распознавания и синтеза речи, изучаются акустические характеристики речевых сигналов, и предпринимаются попытки установления связи между акустическими параметрами и фонетическими признаками речевых сигналов, т.е. попытки понять, как мозг, получив информацию о характере изменения звукового давления во времени, извлекает информацию о смысловом содержании речи. В этом направлении получено уже очень много результатов: количество книг и статей по этим проблемам исчисляется тысячами, в качестве примера могу привести одну из последних книг знаменитого ученого М. Шредера "Computer Speech: Recognition, Compression, Synthesis" (Берлин, 1999 г.).
Однако изучение чисто акустических характеристик речевых сигналов представляет значительную самостоятельную ценность для систем звукозаписи, радиовещания, компьютерной обработки речи
и др., т.е. для всех процессов записи, обработки, передачи и воспроизведения речевых сигналов,
которые принципиально важны для работы звукорежиссера. Поэтому начнем с анализа акустических

характеристик речевых сигналов, а затем попробуем остановиться на их связи с фонетическими признаками, и на существующих в настоящее время теориях слухового восприятия и обработки речи.

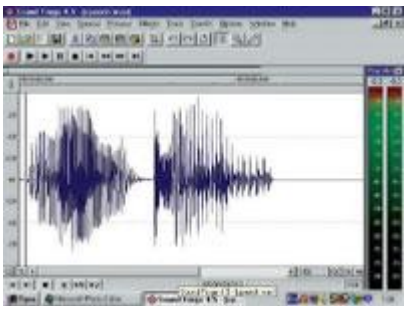

Рис. 1. *Уровнеграмма речевого сигнала*

Анализ акустических характеристик речевого сигнала начинается с записи изменения звукового давления во времени с помощью микрофона – эта зависимость мгновенного значения звукового давления от времени представляется в виде осциллограммы. Обычно в техническим приложениях, в частности при компьютерной обработке, происходит запись усредненного за некоторый отрезок времени уровня звукового давления от времени, эта зависимость называется уровнеграммой. Пример уровнеграммы слова "welcome" показан на рисунке 1.

Вид уровнеграммы существенно зависит от времени и способа усреднения – во всех звуковых программах об этом запрашивается пользователь, (правда, как показывает практика, он не всегда об этом догадывается). Способ усреднения может быть равномерный или экспоненциальный (например, uniform или exponent в программе Sound Forge). Обычно выбирается время усреднения для пиковой уровнеграммы 1…2 мс, для обьективной 15…20 мс, и для субьективной 150…200 мс. В первом случае получается точная запись пиковых значений сигнала; во втором отсутствуют излишние мелкие детали (это время обычно используется при компьютерной обработке речи); наконец, в последнем выбрано время, в течение которого слуховая система опознает тембр.

Если средние значения сигналов сохраняются равными на определенных отрезках времени, то такие сигналы называются стационарными. Звуковые сигналы (речевые и музыкальные) являются сигналами квазислучайными и нестационарными, хотя для речи можно указать приближенно такие отрезки времени (порядка 2…3 мин), при которых речевые сигналы можно считать квазистационарными.

Полученные уровнеграммы позволяют провести статистический, корреляционный и спектральный анализы речевого сигнала, что можно делать с помощью обычных аудиопрограмм, а также с помощью специальных программ, предназначенных именно для речевых сигналов с учетом их специфики: Ultra- sound (Австралия), CSRE (Англия), Viper (Германия), Praat (Голландия), Phonograph (Россия) и др.

Поскольку речевой сигнал, как и музыкальный, представляет собой сигнал квазислучайный, т.е. предсказать его будущие значения можно только с определенной вероятностью, то для анализа его характеристик могут быть применены все известные методы статистического анализа. При этом исследуется распределение во времени следующих величин:

- мгновенных значений и уровней речевого сигнала;
- длительностей непрерывного существования разных уровней;
- длительностей пауз;
- распределение максимальных уровней по частоте;
- распределение текущей и средней мощности;
- спектральной плотности мощности.

Кроме того, могут быть определены такие важные для практики звукозаписи параметры, как динамический диапазон и пик-фактор, вычислено распределение основной фонационной частоты, спектральное распределение формант и др.

Знание статистических характеристик речевых сигналов необходимо для оптимальной организации систем звукового вещания, систем звукозаписи, современных систем сжатия

речевого сигнала и др. Исследование этих характеристик для русской речи было выполнено в работах Фурдуева, Римского-Корсакова, Сапожкова, Белкина, Шитова и др.

Непосредственно из анализа уровнеграмм речевого сигнала прежде всего может быть получена информация о распределении мгновенных значений и уровней звукового сигнала во времени, и длительности их превышения установленного значения. Это позволяет определить динамический диапазон и пик-фактор речевого сигнала, а также установить распределение длительности пауз,
отрезков непрерывных речевых звучаний, распределения текущей и средней мощности сигнала во времени и др.

Если очень коротко остановиться на этих данных, то можно отметить, что распределение плотности вероятности мгновенных значений речевого сигнала, показанное на рисунке 2, носит экспоненциальный характер, и существенно отличается от нормального распределения, которому подчиняется например, джазовая или хоровая музыка. Статистический анализ длительности непрерывного существования разных уровней в речевом сигнале показывает, что наиболее вероятными являются выбросы (пики) длительностью 12…17 мс, из чего следует, что максимальные уровни сигнала достигаются в кратковременные промежутки времени.

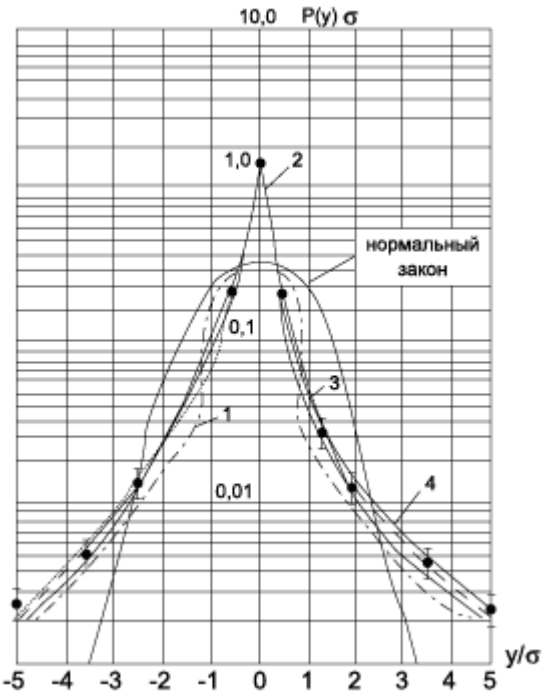

Рис. 2. *Распределение плотности вероятности мгновенных значений речевого сигнала. 1 – дикторский текст; 2, 3, 4 – художественное чтение*

Распределение длительности пауз в речевых сигналах также носит случайный характер, их средняя длительность для речи составляет 0,4 с, а суммарная длительность пауз достигает 5% от времени передачи. Наиболее важная информация, которую позволяет получить анализ уровнеграмм – это определение динамического диапазона речевого сигнала и его пик-фактора.
Динамическим диапазоном звукового сигнала называется разница между его квазимаксимальным и квазиминимальным уровнем D = Lmax – Lmin. Под квазимаксимальным Lmax понимается такой уровень сигнала, длительность пиков выше которого составляет 1% (для речи) и 2% (для музыки) от общей длительности отрезка сигнала. Аналогично определяется квазиминимальный уровень Lmin (относительная длительность составляет 99% и 98%). Значения пик-фактора определяются как разница между квазимаксимальным и средним уровнем сигнала D = Lmax – Lcp.

Значения динамических диапазонов речевых сигналов находится в пределах 35…45 дБ, значения пик-фактора 10…12 дБ.

Условия	Расстояние (см)	Среднее звуковое давление, Па (дБ)	Пиковое значение мощности (мВт)	Пик-фактор (дБ)	Область максимальных уровней (Гц)

Некоторые данные для речевого сигнала по развиваемым уровням звукового давления и мощности приведены в таблице.

Речь телефонная	2,5				
средний уровень		2 (100)	0,24	12	250-500
громкий		4 (106)	4	18	500-1000
тихий		1 (94)	0,025	8	250-500
Разговор	100	0,05 (68)	0,5	10	250-500
Оратор	100	0.1 (74)	2.0	12	250-500

Если пересчитать уровни звукового давления для телефонной речи на расстояние 100 см, то получатся следующие значения: 68, 74, 62 дБ.

Следует отметить ,что для вокальной речи (пения) эти уровни существенно выше, и могут достигать значений 115 дБ на 1 м. В старом итальянском руководстве по подготовке певцов было написано, что если певец может развивать уровень от 110 дБ и выше, то он может петь в "Ла Скала", если ниже 100 дБ, то в камерном ансамбле, если ниже 90 дБ, то не надо петь вообще… Интересно, сколько народу осталось бы петь на эстраде сегодня при таком критерии?

Корреляционный анализ речевого сигнала позволяет рассчитать функцию текущей автокорреляции и установить предел однородности, которые определяются временем, в течение которого функция
автокорреляции достигает некоторого предельного значения, независящего от времени запаздывания. Для речи этот предел составляет 3…5 с.

Спектральный анализ речевого сигнала, как всякого непрерывно изменяющегося во времени акустического сигнала, может быть выполнен на основе записанной уровнеграммы с помощью преобразования Фурье. В любом музыкальном редакторе предусмотрена операция быстрого преобразования Фурье (БПФ, FFT), позволяющая из выделенного отрезка уровнеграммы рассчитать его спектр.

Анализ спектров речевых сигналов позволяет установить форму огибающей и выделить области формантных частот. Поскольку место и ширина формантных областей принципиально важны для распознавания речи, то для точного определения формантных полос в речевом сигнале созданы специальные программы на основе метода линейного предсказания или кепстрального анализа, позволяющие производить их автоматическое распознавание.

Кроме того, поскольку интонация речевого высказывания определяется изменением частоты фонации, то выделение основной частоты фонации из записанных уровнеграмм и характер ее зависимости от времени имеют принципиально важное значение.

Для интегральной оценки свойств речевого сигнала может быть рассчитан спектр мощности и построено распределение спектральной плотности мощности, которая для речевого сигнала показана на рисунке 3, что позволяет установить, что основная энергия речевого сигнала (В) сосредоточена в полосе 250…1000 Гц, спад в сторону высоких частот происходит со скоростью 7 дБ/окт после 500 Гц.

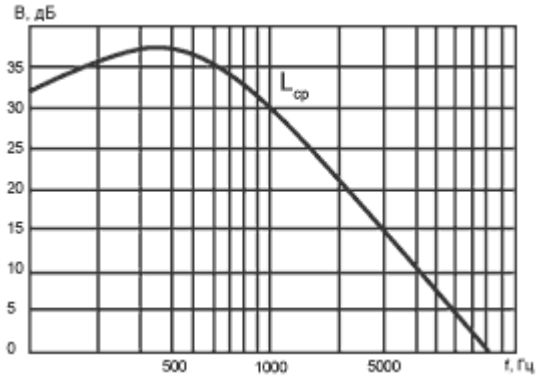

***Рис. 3**. Спектральное распределение средней мощности речевого сигнала*

Анализ спектров дает возможность построить очень важную для практики звукозаписи кривую распределения амплитудного состава речи. Пример для диапазона 1000…1400 Гц показан на рисунке 4, (для других диапазонов распределения аналогичные). Кривая распределения показывает, что более 80% в речевом потоке составляют амплитуды с уровнем 45 дБ, и только менее10% амплитуды с уровнями 70 дБ и выше. Это значит, что при обработке речевых фонограмм стремление "вычистить шумы" может привести к потере значительной части информации, поскольку низкие уровни амплитуд связаны в основном с согласными звуками, а они являются носителями основной смысловой нагрузки в речи.

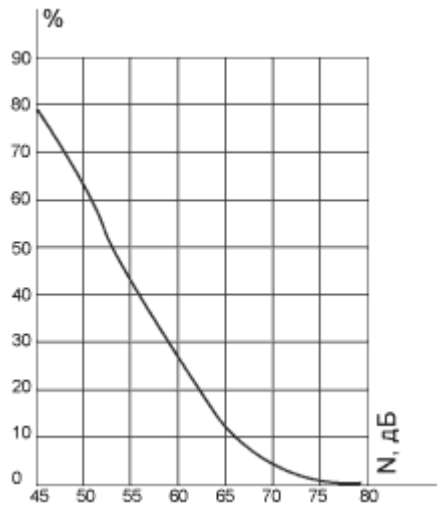

***Рис. 4**. Амплитудный состав речи в полосе 1000 - 1400 Гц*

Кроме одномерных спектров (амплитуда-частота), современные алгоритмы позволяют построить для любого речевого сигнала его трехмерные (кумулятивные) спектры (например, 3D-Frequency Analysis в редакторах Wave-Lab и др.), где по одной оси отложено время, по другой частота, по третьей – амплитуда. (Рисунок 5). Такие спектры позволяют получить значительно больше информации не только о спектральном составе сигнала, но и характере изменения его во времени. Трехмерные спектры широко используются в практике изучения различных акустических сигналов, однако для анализа речевых сигналов наибольшее распространение имеют трехмерные спектры особой формы-спектрограммы.

В 1940 году в лаборатории Bell Lab (США) был построен прибор, получивший название "спектрограф видимой речи", который позволял представить спектр речи в трехмерной форме, только построенной несколько иначе, чем обычный трехмерный спектр. Это своего рода "вид сверху" на трехмерный спектр: по оси абсцисс отложено время, по оси ординат – частота, а амплитуда показана интенсивностью цвета (чем интенсивнее, тем больше амплитуда). На рисунке 6 показан пример спектрограммы того же речевого сигнала, 3D-спектр которого дан на рисунке 5.

Спектрограммы могут быть узкополосные, широкополосные и слуховые. Выбор числа семплов, т.е. выбор длительности отрезка анализируемого сигнала, определяет точность развертывания по частоте (т.е. расстояние между частотами). Невозможно обеспечить одновременно "хорошее" развертывание и по частоте и по времени, поскольку они связаны некоторым соотношением Df•Dt = const,
(по аналогии с квантовой механикой называемым "принципом неопределенности"). Чем выше точность по частоте, тем хуже развертывание по времени, и наоборот. Поэтому точность развертки по частоте зависит в обратной пропорции от длительности временного окна при преобразовании Фурье (например, при ширине развертки 100 Гц развертывание по времени будет 1/100 = 10 мс).

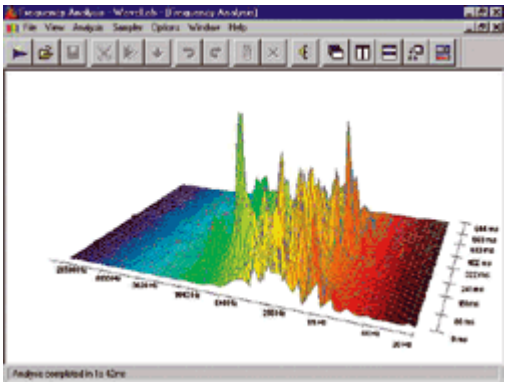

Рис. 5. *Трехмерный (кумулятивный) спектр речевого сигнала*

В практике анализа речевых сигналов применяется два вида спектрограмм: широкополосные и узкополосные (рисунки 7а, 7б). В узкополосных спектрограммах используется частота развертки 45 Гц, это ниже, чем самые низкие фонационные частоты в голосе, что позволяет при такой точной развертке отчетливо увидеть вдоль вертикальной оси гармоники голосового источника.

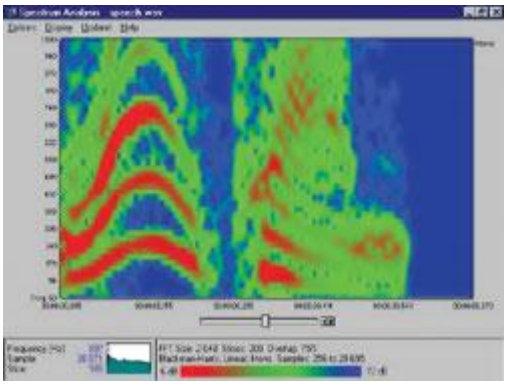

Рис. 6. *Спектрограмма речевого сигнала*

Как было сказано в предыдущих статьях, речевой сигнал – это результат "свертки" (умножения) звукового сигнала, создаваемого голосовым источником, например, за счет модуляции воздуха при колебаниях голосовых связок, и огибающей, за счет резонансных свойств голосового тракта (этим и определяется его формантная структура.

На широкополосных спектрограммах, обычно с частотой развертывания 300 Гц, отчетливо видны вертикальные полосы вдоль оси времени, связанные с появлением отдельных импульсов воздушного давления при колебаниях голосовых связок, и сильно подчеркнуты темные горизонтальные полосы, соответствующие формантам. Поэтому, в зависимости от целей, которые ставятся при анализе речевого сигнала, используются или широкополосные спектрограммы (выделяются отдельные импульсы воздуха, подчеркнуты форманты), или узкополосные, где выделяются обертоны голосового источника. При этом можно проследить изменение основной частоты фонации во времени, что имеет большое значение при оценке мелодического рисунка речи, как было отмечено выше. Кроме того, полученные значения спектров позволяют оценить распределение энергии во времени.

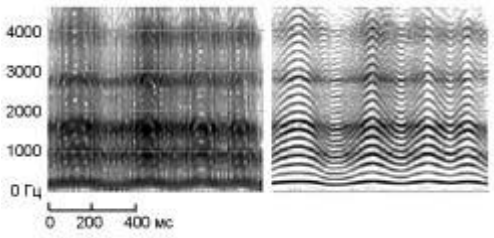

Рис. 7. *а) Широкополосная спектрограмма; б) узкополосная спектрограмма*

Однако, ни широкополосная, ни узкополосная спектрограммы не учитывают специфику спектрального анализа сигнала, который производится во внутреннем отделе периферической слуховой системы на базилярной мембране. Поэтому в последние годы с учетом новейших результатов в психоакустике была разработана методика построения "слуховых"

спектрограмм. При построении этих спектрограмм используются фильтры с различными полосами пропускания, ширина которых соответствует ширине "критических полос" слуха (или ширине слуховых фильтров при спектральном анализе звуков на базилярной мембране).

Ширина критических полос зависит от частоты, эта зависимость примерно соответствует ширине третьоктавных полос. В такой спектрограмме на низких частотах (первые 4…5 критических полос) происходит узкополосная обработка сигнала по частоте. На высоких частотах критические полосы становятся значительно шире, это соответствует широкополосной спектрограмме, т.е. идет очень точное развертывание по времени.

Таким образом, слуховая спектрограмма значительно точнее отражает восприятие и обработку
речевого сигнала в слуховой системе: на низких частотах основное внимание концентрируется на отдельных гармониках, на высоких производится интегральная оценка гармоник, но зато точно
отслеживается динамика изменения их огибающей во времени – аналогично тому, как это происходит при оценке высоты тона.

В итоге, в низкочастотной области слух оценивает значение основной частоты фонации и ее первых обертонов, и по ним определяет высоту голоса; в верхней части слух точно оценивает изменение огибающей во времени, что позволяет ему выделить формантную картину, которая служит базовой информацией для верхних отделов мозга при определении фонетического значения отдельных фонем, слогов и др.

Таким образом, при анализе акустических параметров речевого сигнала в современных специализированных программах оцениваются следующие характеристики:
- уровнеграмма и все связанные с ней параметры
(динамический диапазон, распределение мгновенных значений сигнала, текущая мощность и др.);
- одномерный спектр (распределение

формантных областей);
- трехмерный спектр (изменение формы огибающей во времени);
- спектрограммы (широкополосные, узкополосные, слуховые), из которых могут быть получены такие характеристики, как изменение основной фонационной частоты во времени, изменение формантных областей, распределение гармоник голосового источника, временная структура импульсов звукового давления и др.

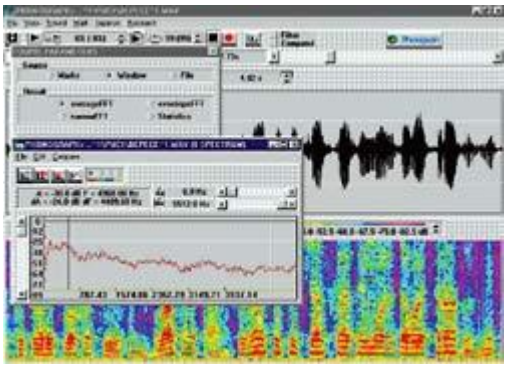

Рис. 8. *Пример анализа речевого сигнала*

Кроме того, в ряде программ предусмотрена операция расчета нелинейной маскировки составляющих
речевого сигнала, удаление неслышимых компонент
расчет распределения формантных полос с учетом их ширины и добротности. др. Общая картина анализа речевого сигнала, обычно производимая в современных компьютерных программах, показана на рисунке 8.